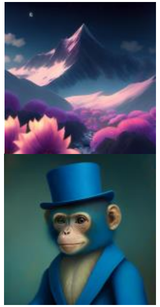
Di[M]O: Distilling Masked Diffusion Models into One-step Generator

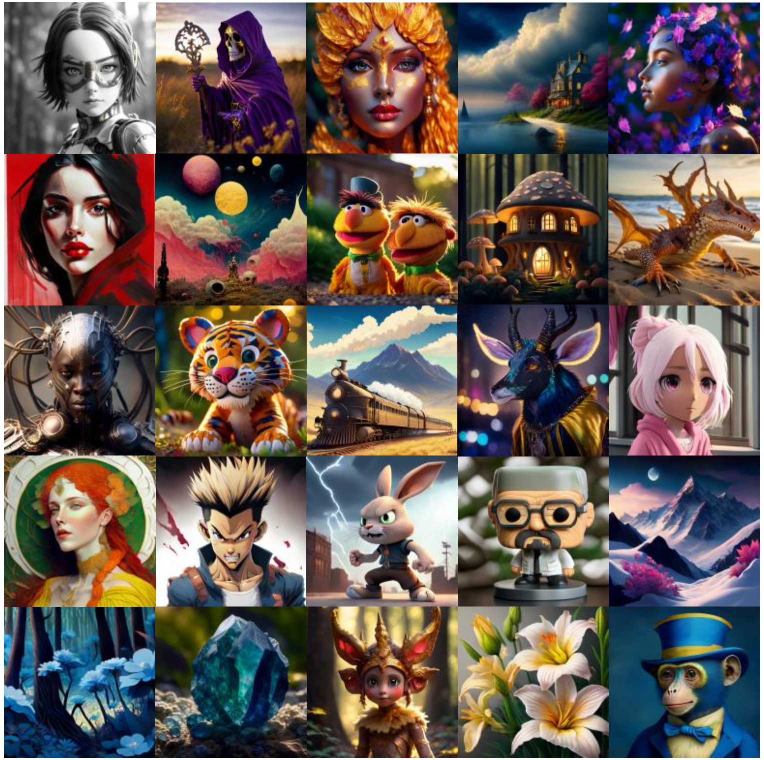
Yuanzhi Zhu, Xi Wang, Stéphane Lathuilière, Vicky Kalogeiton

LIX, École Polytechnique, CNRS, IPP, Inria, Univ. Grenoble Alpes, CNRS, LJK

Runtime Comparison



10.5 s/image



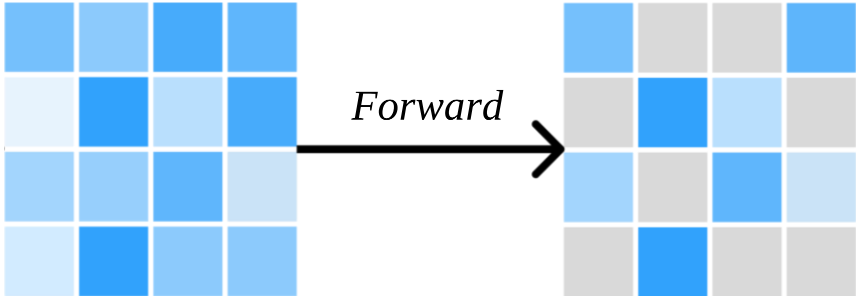
0.80 s/image

Content

- Preliminaries
- Methods
- Results

Mask Diffusion Models

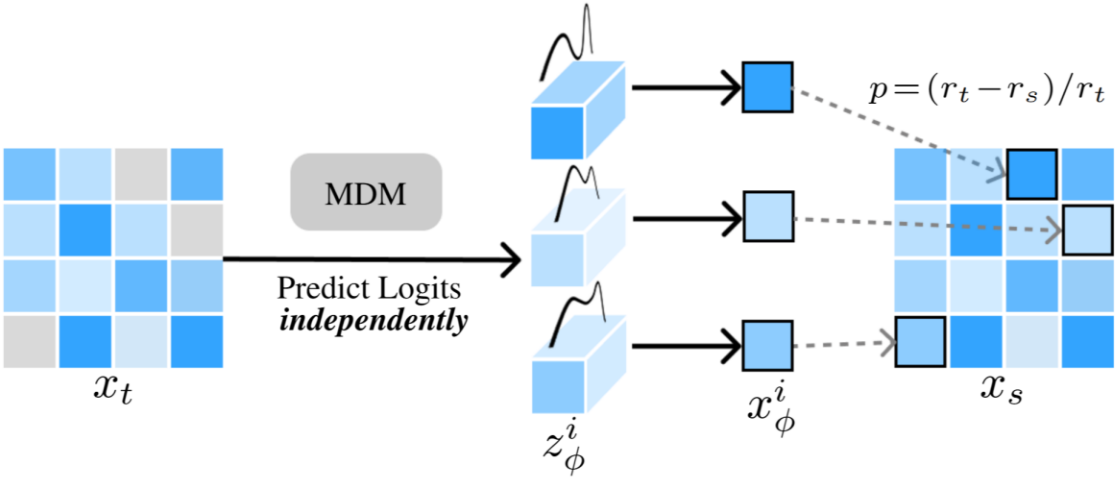
Forward Process: mask *independently* each token at position *i*



$$q_{t|0}(x_t|x_0) = \prod_{i=0}^{L-1} \mathrm{Cat}(x_t^i; (1-r_t)\delta(x_0^i) + r_t\delta([\mathtt{M}]))$$

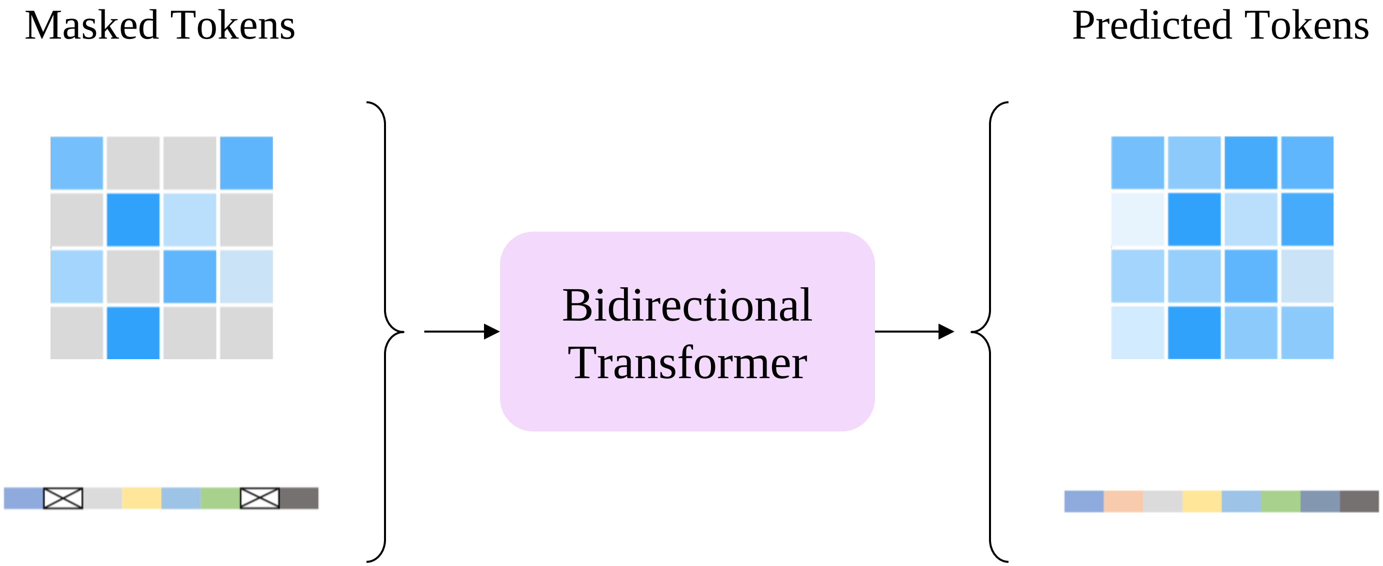
Mask Diffusion Models

Reverse Process: fill mask conditional *independently*



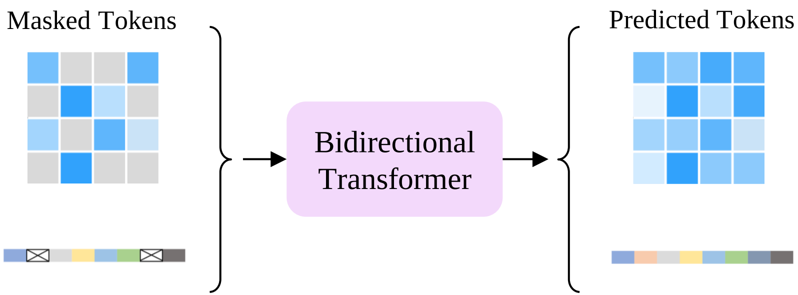
Predicted Probability: $p_\phi(x_0^i|x_t) := \text{softmax}\left(z_\phi^i(x_t)/\tau\right)$

Mask Diffusion Models

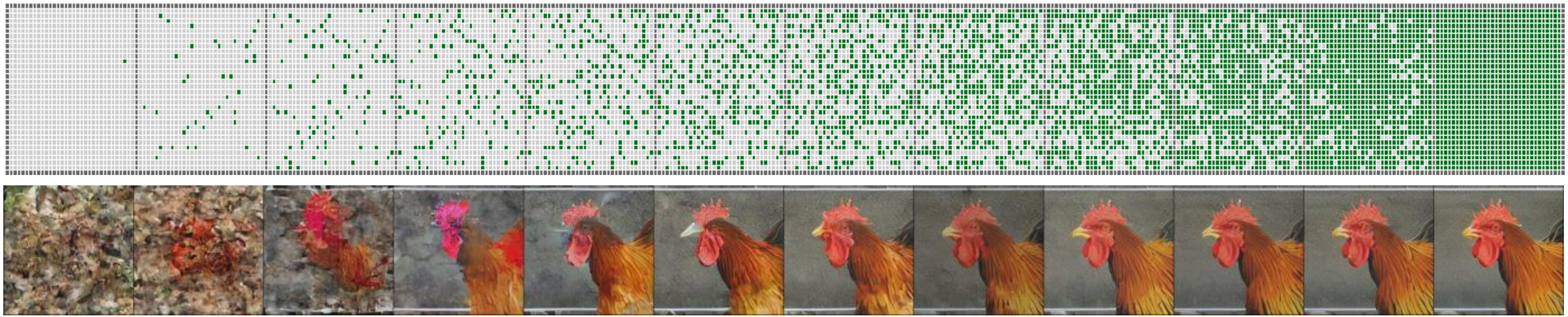


$$\mathcal{L}_{\text{MDM}} = \mathbb{E}_{x_0,t}\left[\left(\mathbb{E}_{q_{t|0}}[-\log p_{0|t}(x_0|x_t,\phi)]\right)\right]$$

Mask Diffusion Models

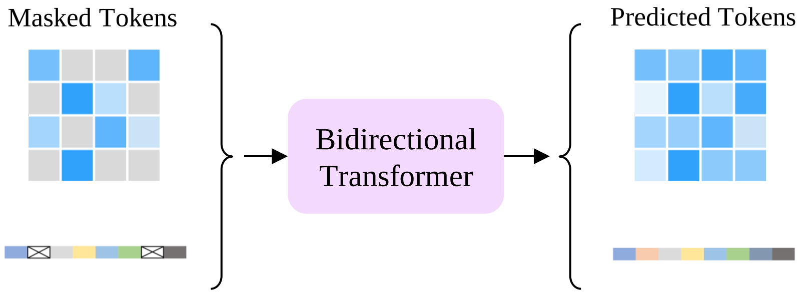


Sampling Process



[2202.04200] MaskGIT: Masked Generative Image Transformer

Promise of MDM



$$\mathcal{L}_{\text{MDM}} = \mathbb{E}_{x_0,t}\left[\left(\mathbb{E}_{q_{t|0}}[-\log p_{0|t}(x_0|x_t,\phi)]\right)\right]$$

Theorem: Minimizing the *expected* cross-entropy *loss* is equivalent to maximizing the upper bound on the negative log-likelihood, i.e.:

$$-\mathbb{E}_{p_{\text{data}}(x_0)}[\log p_\theta(x_0)] \leq \mathcal{L}_{\text{MDM}}(\theta)$$

MDM loss (maximize likelihood) + Transformer (less inductive bias) + Data → *Scalability* ?

[2406.03736] Your Absorbing Discrete Diffusion Secretly Models the Conditional Distributions of Clean Data

Limitations of MDM

- Incorrect joint distribution for multi-token prediction ☹
- Need sufficient steps for better performance ☹

Ideas in Inference-time Scaling can Benefit Generative Pre-training Algorithms

Jiaming Song, Linqi Zhou

Luma AI

Multi-token prediction (MTP) is of great interest to the language modeling community because of its potential to achieve faster inference [GIR+24], which allows efficient inference-time scaling. However, the current multi-token prediction models often predict the softmax values of multiple tokens in parallel, which is a naïve conditional independence assumption (*i.e.*, naïve Bayes). We argue that this inference design greatly limits the capacity of the model distribution and more efforts should be spent resolving this fundamental issue.

Theoretical Benefit and Limitation of Diffusion Language Model

Guhao Feng [*1] **Yihan Geng** [*1] **Jian Guan** [2] **Wei Wu** [2] **Liwei Wang** [1] **Di He** [1]

such as a reasoning chain–we show that the required sampling steps must scale linearly with sequence length to obtain "correct" sequences, thereby eliminating MDM's efficiency advantage over autoregressive models. Our analysis estab-

[2503.07154] Ideas in Inference-time Scaling can Benefit Generative Pre-training Algorithms

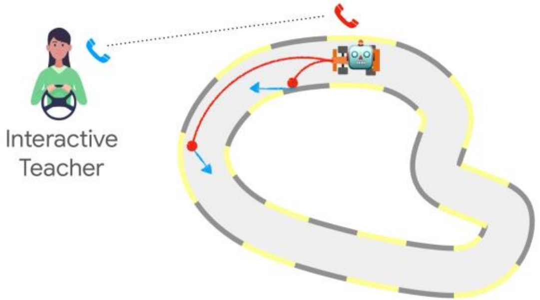
[2502.09622] Theoretical Benefit and Limitation of Diffusion Language Model

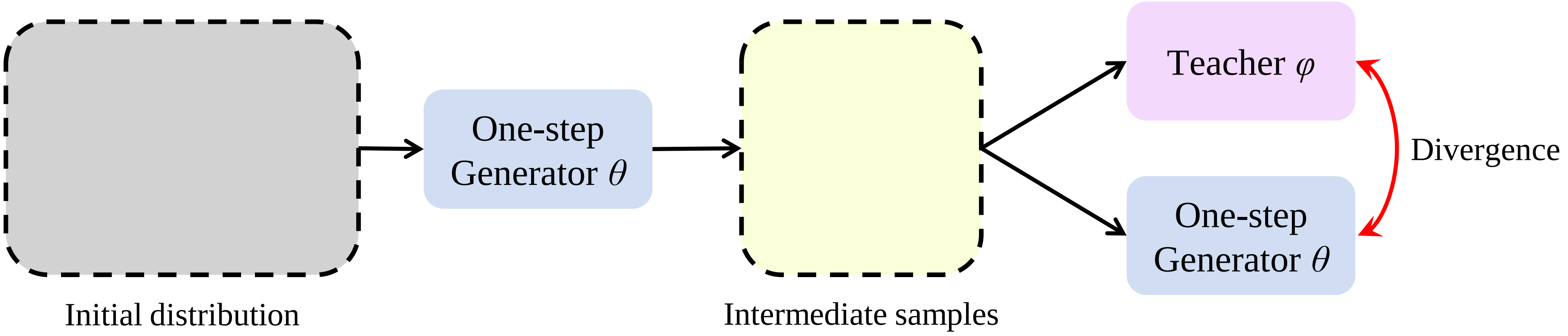
Content

- Preliminaries
- Methods
- Results

Main Idea: Distillation in an *On-Policy* fashion

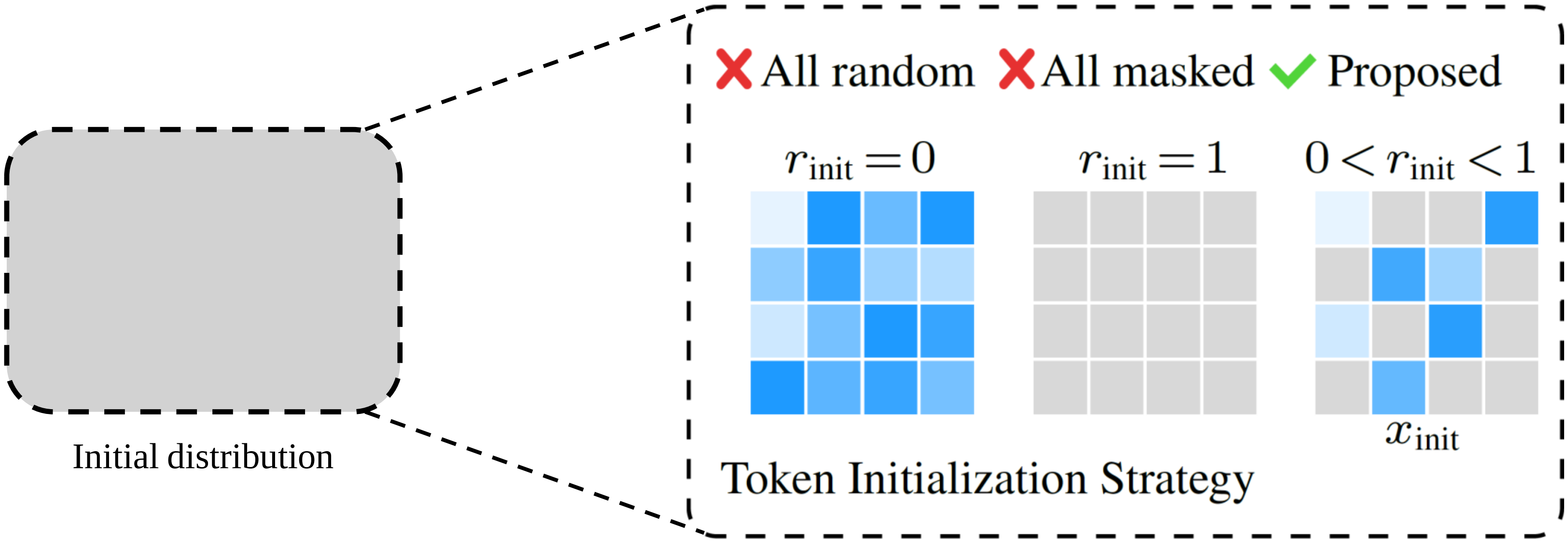
Match teacher and generator for all possible intermediate samples





$$\mathcal{L}_{\mathrm{Di[M]O}}(\theta) := \mathbb{E}_{x_{\mathrm{init}},t}\left[w(t)\left(\mathbb{E}_{q_{t|0}}[D(p_\phi||p_\theta)(\tilde{x}_t)]\right)\right]$$

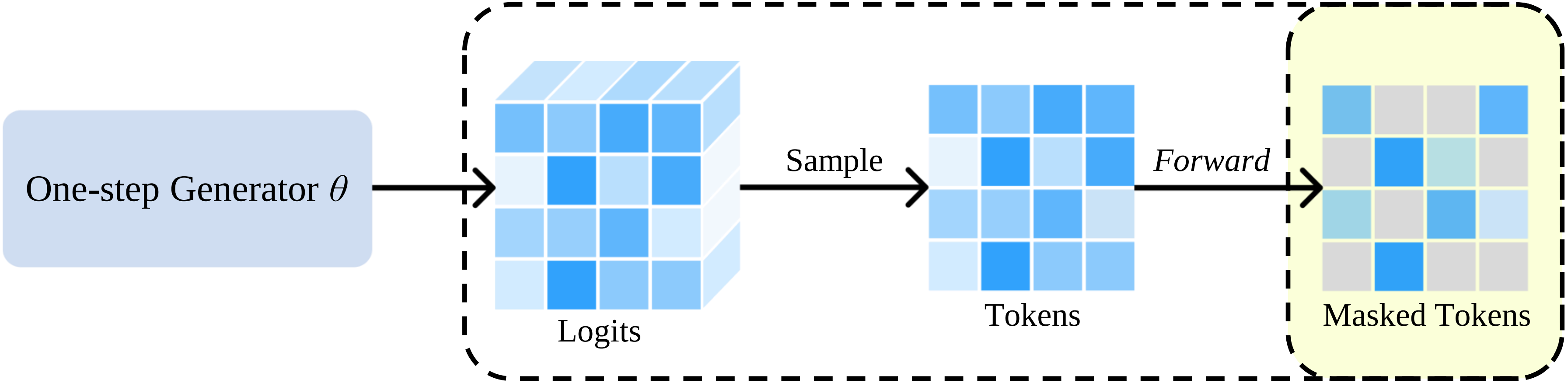
Initial Distribution



Close to the input distribution *(masked token sequences)* used to train the teacher model

Intermediate Samples

Via *forward mask diffusion process*



Consider loss on each intermediate state $\tilde{x}_t$

$$\mathcal{L}_{\mathrm{Di[M]O}}(\theta) := \mathbb{E}_{x_{\mathrm{init}},t}\left[w(t)\left(\mathbb{E}_{q_{t|0}}\left[\boxed{D(p_\phi \| p_\theta)(\tilde{x}_t)}\right]\right)\right]$$

Further decompose into ***Token-level*** Divergence

$$\boxed{D(p_\phi \| p_\theta)(\tilde{x}_t) := \frac{1}{L_M} \sum_{\substack{i=1 \\ \tilde{x}_t^i = [\mathtt{M}]}}^{L} D\left(p_\phi(x_0^i | \tilde{x}_t) \| p_\theta(x_0^i | \tilde{x}_t)\right)}$$

Gradient of Divergence:

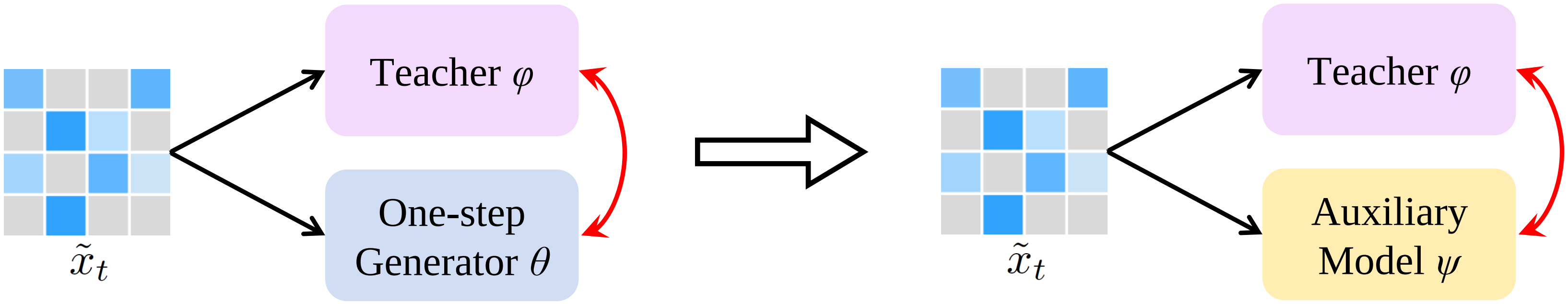
$$\nabla_\theta D\big((p_\phi || p_\theta)(\tilde{x}_t)\big) = \nabla_{z_\theta} D(p_\phi || p_\theta)(\tilde{x}_t) \frac{\mathrm{d} z_\theta(\tilde{x}_t)}{\mathrm{d}\theta}$$

Intractable terms

$$p_\theta(x_0^i | \tilde{x}_t) = \mathrm{softmax}\big(z_\theta(\tilde{x}_t) \big)$$

Approximation: $p_\theta(x_0^i|\tilde{x}_t)$

Introduce an ***auxiliary model*** to approximate the student output on intermediate samples



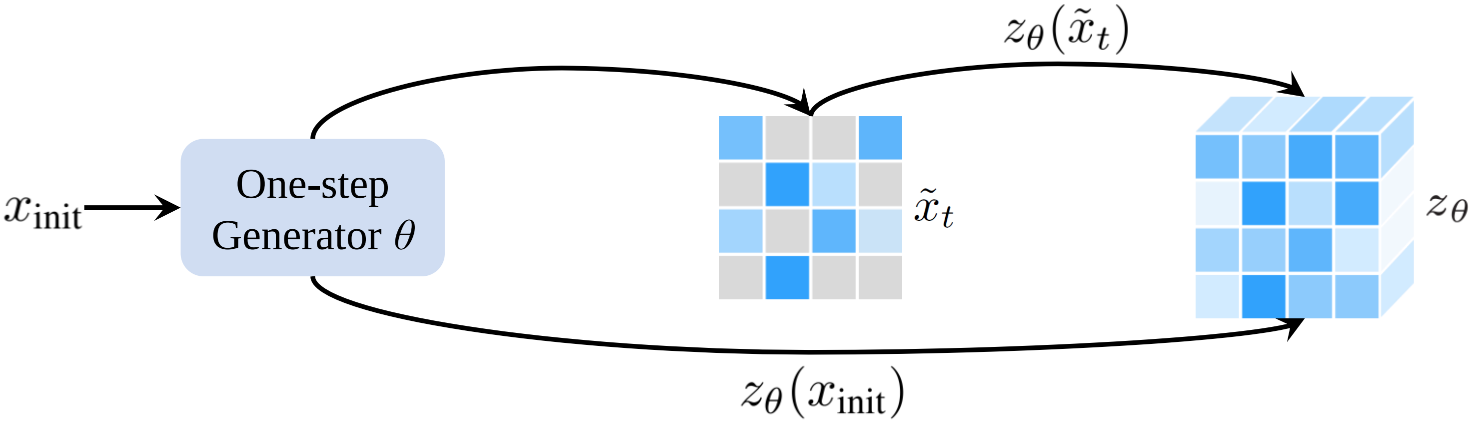
$$p_\psi(x_0^i|\tilde{x}_t) \approx p_\theta(x_0^i|\tilde{x}_t)$$

$$\nabla_{z_\theta} D(p_\phi||p_\theta)(\tilde{x}_t) \Longrightarrow \nabla_{z_\psi} D(p_\phi||p_\psi)(\tilde{x}_t)$$

Approximation: $z_\theta(\tilde{x}_t)$

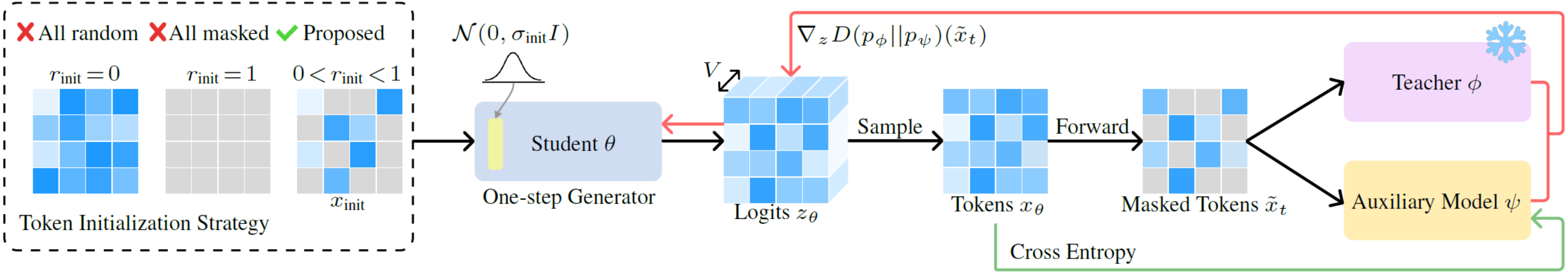
We approximate the output logits based on the *consistency assumption*

Also, we require this term as effective gradient for the generator



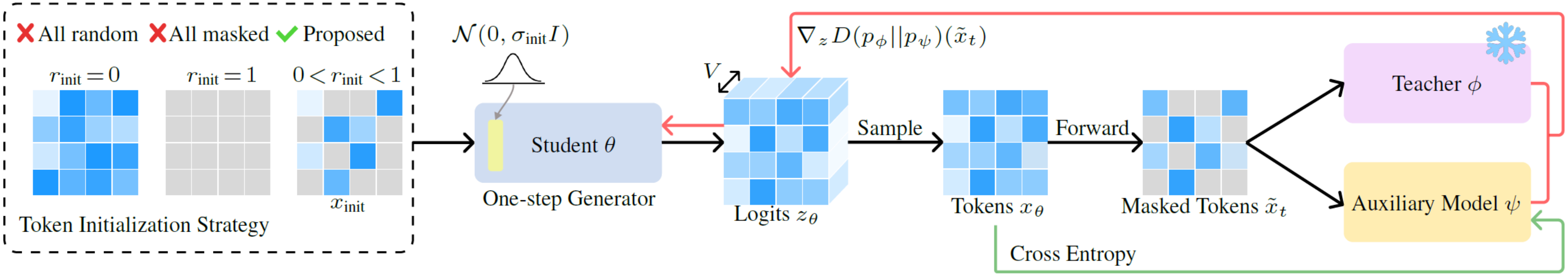
$$\frac{\mathrm{d}z_\theta(\tilde{x}_t)}{\mathrm{d}\theta} \Rightarrow \frac{\mathrm{d}z_\theta(x_{\text{init}})}{\mathrm{d}\theta}$$

Overview of the Method

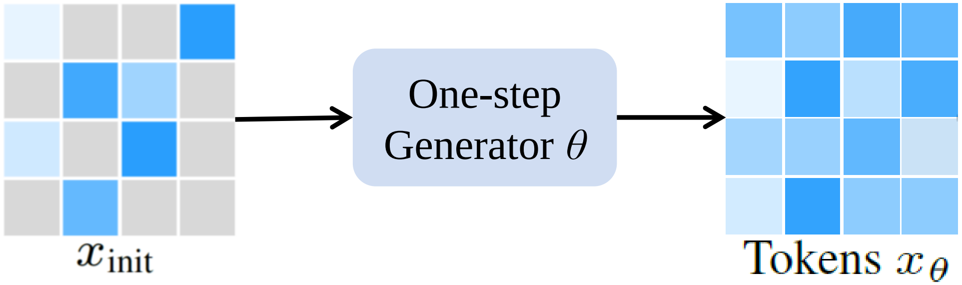


$$\nabla_\theta \mathcal{L}_{\text{Di[M]O}} \approx \mathbb{E}_{x_{\text{init}}, t}\left[w(t) \left(\mathbb{E}_{q_{t|0}} \left[\nabla_{z_\psi} D(p_\phi || p_\psi)(\tilde{x}_t) \frac{\mathrm{d} z_\theta(x_{\text{init}})}{\mathrm{d}\theta} \right] \right) \right]$$

Overview of the Method



- From random initial states to samples from *correct joint distribution* ☺
- Greatly reduce the sampling steps → One step generation ☺



Content

- Preliminaries
- Methods
- Results

Visual Results

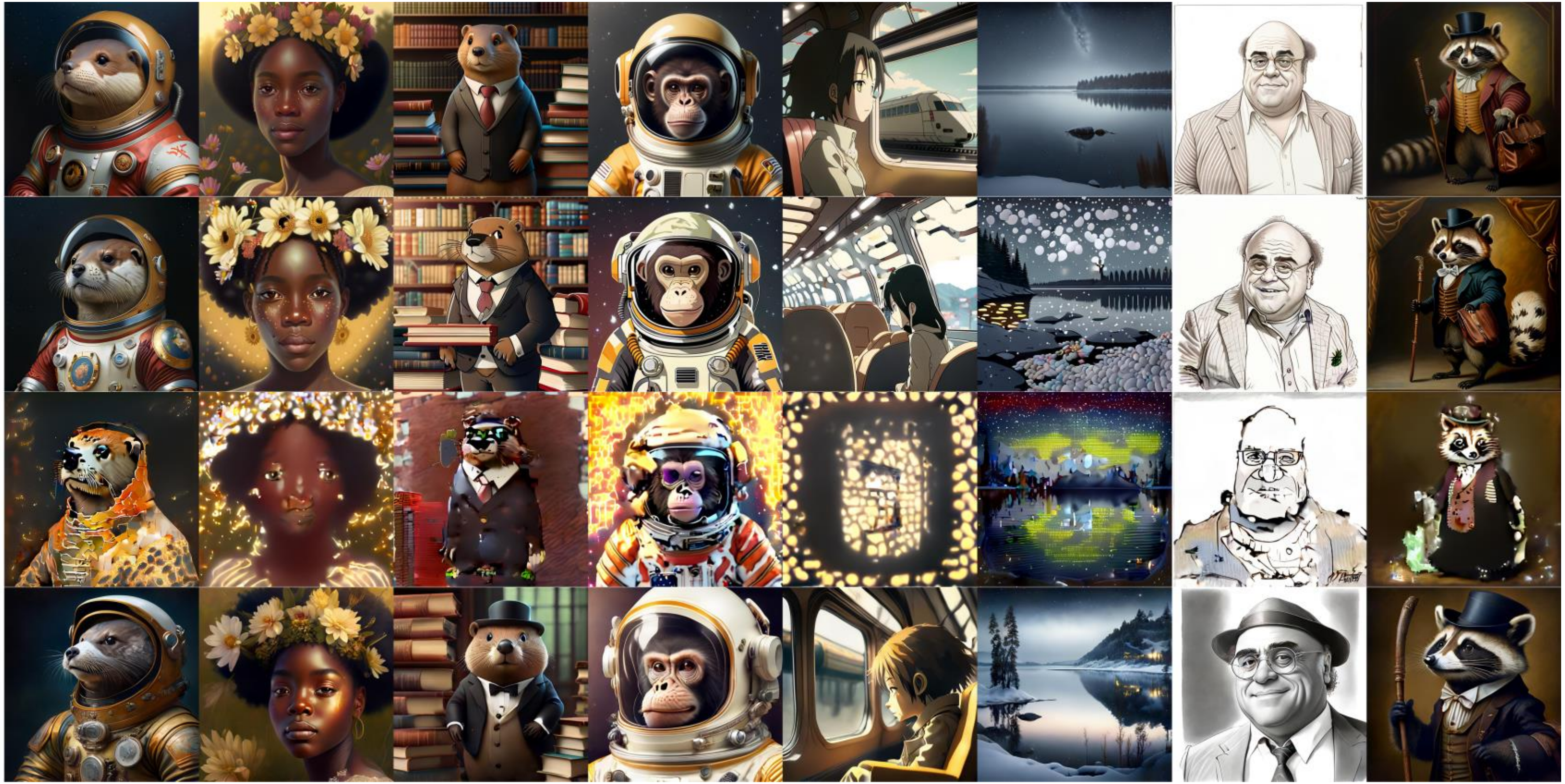
Visual Results

64 Steps

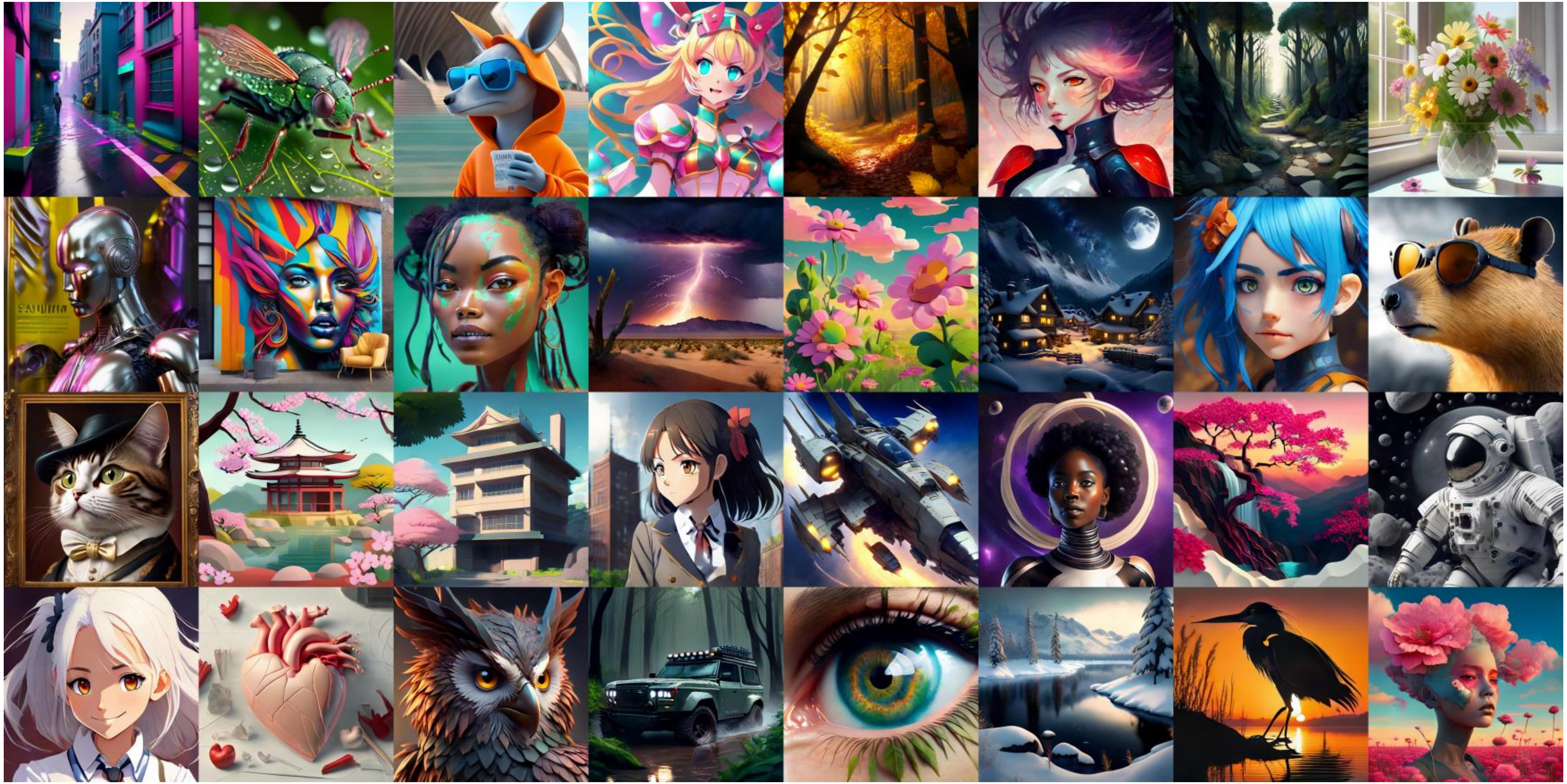
16 Steps

4 Steps

1 Step (ours)



Visual Results



Quantitative Results (MaskGit Teacher)

ImageNet	Method	Step (↓)	FID (↓)	IS (↑)	Precision (↑)	Recall (↑)	Density (↑)	Coverage (↑)
Teacher	MaskGit [7]	16	6.60	224.07	0.831	0.402	1.246	0.977
	MaskGit [7]	8	6.66	221.57	0.827	0.397	1.233	0.974
	MaskGit [7]	4	10.73	192.29	0.748	0.313	1.011	0.920
	MaskGit [7]	2	91.35	13.37	0.178	0.164	0.091	0.122
Sampler	θ-trapezoidal [81]*	64	6.7	-	-	-	-	-
	θ-trapezoidal [81]*	32	7.1	-	-	-	-	-
Distillation	di4c [31]	4	6.79	209.2	-	-	-	-
	di4c-d [31]	4	6.57	213.6	-	-	-	-
	Di[M]O	**1**	6.91	214.0	0.828	0.377	1.255	0.967

Quantitative Results (Meissonic Teacher)

HPS v2.0						
Model	**Step**	**Anim.**	**Concept-art**	**Painting**	**Photo**	**Averaged**
Latent Diffusion [83]	25	25.73	25.15	25.25	26.97	25.78
DALL·E 2 [79]	-	27.34	26.54	26.68	27.24	26.95
Stable Diffusion v1.4 [83]	50	27.26	26.61	26.66	27.27	26.95
Stable Diffusion v2.0 [83]	50	27.48	26.89	26.86	27.46	27.17
DeepFloyd-XL [19]	25	27.64	26.83	26.86	27.75	27.27
SDXL Base 1.0 [76]	50	28.88	27.88	27.92	28.31	28.25
SDXL Refiner 1.0 [76]	50	28.93	27.89	27.90	28.38	28.27
InstaFlow [55]	1	25.98	25.79	25.93	26.32	26.01
SD Turbo [90]	1	27.98	27.59	27.16	27.19	27.48
SwiftBrush v2 [18]	1	27.25	27.62	26.86	26.77	27.15
Meissonic [4]	48	**29.57**	**28.58**	**28.72**	**28.45**	**28.83**
	32	29.18	28.32	28.28	27.96	28.44
	16	28.61	27.82	27.84	27.32	27.90
	8	25.62	26.49	26.67	27.07	26.46
	4	25.01	24.95	24.87	23.80	24.66
	2	23.06	23.28	23.22	22.38	22.98
Di[M]O	1	28.64	27.91	27.99	27.92	28.11

GenEval								
Model	**Step**	**Overall**	**Objects**		**Counting**	**Colors**	**Position**	**Color Attribution**
			Single	**Two**				
SD v1.5 [83]	50	0.43	0.97	0.38	0.35	0.76	0.04	0.06
SD v2.1 [83]	50	0.50	0.98	0.51	0.44	0.85	0.07	0.17
DALL-E2 [79]	-	0.52	0.94	0.66	**0.49**	0.77	0.10	0.19
SDXL [76]	50	**0.55**	0.98	**0.74**	0.39	0.85	**0.15**	**0.23**
Meissonic [4]	48	0.54	**0.99**	0.66	0.42	**0.86**	0.10	0.22
	32	0.46	0.92	0.53	0.33	0.80	0.08	0.13
	16	0.37	0.82	0.39	0.20	0.70	0.05	0.08
	8	0.20	0.58	0.12	0.05	0.40	0.02	0.04
	4	0.09	0.31	0.02	0.01	0.18	0.01	0.01
	2	0.03	0.14	0.01	0.00	0.05	0.00	0.00
Di[M]O	1	0.43	0.91	0.53	0.22	0.75	0.07	0.11

Thank You!